ABSTRACT

A method and apparatus for transforming a web page that contains main content and auxiliary data. The web page is converted into a string containing multiple first values and multiple second values. The first values correspond to formatting code segments within the web page and the second values correspond to text segments within the web page. Further, a low-pass filter is applied to the string containing multiple first values and multiple second values, and the output of the low-pass filter is used to determine the location of the main content within the web page.